1. Kostarev V.S., Klimova V.A., and Tashlykov O.L. Simulation of natural cooling modes of containers with radioactive wastes: AIP Conference Proceedings 2015, 020044 (2018)
2. Костарев В.С., Климова В.А., Ташлыков О.Л. Моделирование режимов пассивного охлаждения контейнеров с радиоактивными отходами // Ядерные технологии: от исследований к внедрению. Сборник материалов научно-практической конференции. Нижний Новгород: Нижегородский государственный технический университет им. Р.Е. Алексеева, 2018. С.14-15
3. Елистратов В.В., Грилихес В.А., Аронова Е.С. Солнечные энергоустановки. Оценка поступления солнечного излучения. — СПб.: Издательство Политехнического университета, 2009.

ПРИМЕНЕНИЕ ПАРАМЕТРОВ МЕРЫ ХАОСА И ПОРЯДКА ДЛЯ КЛАССИФИКАЦИИ РУССКОЯЗЫЧНЫХ ТЕКСТОВ

Филимонов В.В.[1*], Живодёров А.А.[1,2], Крамаренко А.А.[1]

[1)] Уральский федеральный университет имени первого Президента России Б.Н. Ельцина, г. Екатеринбург, Россия
[2)] Центральная научная библиотека УрО РАН, г. Екатеринбург, Россия
*E-mail: bukva32@yandex.ru

APPLICATION OF THE PARAMETERS OF MEASURES OF ORDER AND DISORDER FOR CLASSIFICATIONS OF RUSSIAN-LANGUAGE TEXTS

Filimonov V.V.[1*], Zhivodyorov A.A.[1,2], Kramarenko A.A.[1]

[1)] Ural Federal University, Ekaterinburg, Russia
[2)]Central scientific library URAN, Ekaterinburg, Russia

The paper considers the problem of analysis and classification of Russian-language texts using mathematical models. Authors research applicability of the parameters of measures of order and disorder: R-function and D-function.

В настоящее время для анализа текстов широко используются компьютерные методы, основанные на математических моделях. Среди них применение цепей Маркова, использование ранговых распределений и так далее.

Настоящая работа посвящена построению математической модели текста. Используемая методика предполагает численное выражение отдельных параметров, служащих атрибутами текста, и может быть использована для поиска текстов необходимого жанра и стиля, установления авторства и оценки юзабилити текста.

В наших работах использовались методы математической статистики и модель случайных блужданий, при помощи которых был получен ряд параметров текста: величина статистики χ^2, «коэффициент диффузии» текста, частоты появления в тексте отдельных гласных букв, и другие.

В предыдущей работе нами был получен компактный и достаточный набор параметров, дающий правильную классификацию текстов с вероятностью более 80%. Набор включает в себя статистику χ^2, «коэффициент диффузии», степень сжимаемости текста архиватором, а также частоты букв «О» и «Э».

Для повышения качества классификации текстов мы предлагаем дополнительно использовать параметры меры хаоса и порядка в дискретных системах, предложенные В.Б. Вяткиным в работе «Хаос и порядок дискретных систем в свете синергетической теории информации». Такими параметрами являются так называемые R- и D-функции.

R-функция, представляющая собой отношение аддитивной негэнтропии к энтропии отражения, выражает соотношение порядка и хаоса в системе (1).

$$R = \frac{I_\Sigma}{S} \tag{1}$$

D-функция, представляющая собой произведение аддитивной негэнтропии и энтропии отражения, выражает «степень развитости системы», принимая нулевые значения в случае абсолютного хаоса и абсолютного порядка (2).

$$D = I_\Sigma S \tag{2}$$

Величины I_Σ и S вычисляются следующим образом:

$$\mathrm{I}_\Sigma = \sum_{i=1}^{N} \frac{m_i}{M} \log_2 m_i \tag{3}$$

$$\mathrm{S} = -\sum_{i=1}^{N} \frac{m_i}{M} \log_2 \frac{m_i}{M} \tag{4}$$

где M – общее количество элементов в составе системы, N – число частей системы, m_i– количество элементов в i-й части.

Настоящая работа посвящена исследованию пригодности указанных параметров для решения задачи классификации русскоязычных текстов.

1. Вяткин В.Б. Научный журнал КубГАУ, №47(3), 2009, [Электронный ресурс]. — Режим доступа: http://ej.kubagro.ru/2009/03/pdf/08.pdf. (дата обращения: 15.12.2018)